



**University of
Zurich**^{UZH}

**Zurich Open Repository and
Archive**

University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2018

Young GI angle: Planning the statistical analysis of the results

Kirchgesner, Julien ; Misselwitz, Benjamin

DOI: <https://doi.org/10.1177/0391560318813777>

Posted at the Zurich Open Repository and Archive, University of Zurich

ZORA URL: <https://doi.org/10.5167/uzh-161034>

Journal Article

Published Version

Originally published at:

Kirchgesner, Julien; Misselwitz, Benjamin (2018). Young GI angle: Planning the statistical analysis of the results. United European Gastroenterology Journal, 6(10):1598-1600.

DOI: <https://doi.org/10.1177/0391560318813777>

Young GI angle: Planning the statistical analysis of the results

United European Gastroenterology Journal
2018, Vol. 6(10) 1598–1600
© Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0391560318813777
journals.sagepub.com/home/ueg



Introduction

Conducting a research project is one of the first key steps in a young person's gastrointestinal (GI) career. The success of the project will impact on her or his interest in research and motivation them to embrace an academic career. An appropriate methodology includes a statistical analysis plan, which is often seen as the most complex part of the process, which requires efficient data collection and active collaboration with the methodologist–statistician. Several tips can ease this process and increase the productivity of the trainee's work. Here, we will discuss some general guidelines that will help to plan the statistical analysis. We will frequently use colonoscopy screening as an illustrative example.

Analysis plan

From a hunch to a research hypothesis and a clear main objective

A project typically starts with an observation, a question or a hunch. The first challenge is to transform such a vague idea into a clear main objective. As a start, it is worth explaining the question in plain, non-medical language to a colleague. A good hunch can be expressed as a **hypothesis**, which can be confirmed or refuted. Major inconsistencies frequently surface at this point. Science should be viewed as a conversation where an idea, with its reasoning and calculated uncertainty, will be communicated to other scientists.

One should gradually progress from a vague idea to a specific **main objective** expressed in medical language using defined terms (i.e. 'assess the decrease in colorectal cancer (CRC) mortality after a single screening colonoscopy at 50 years compared to no screening').

Analysis plan content

The analysis plan should be fixed before starting data collection. Investing intellectual effort at the beginning will save time during data collection and later. In addition, a pre-specified analysis plan greatly increases the credibility of study results, guarding against data dredging. A study protocol and analysis plan can be

registered at various places (e.g. clinicaltrials.gov for randomized controlled trials and observational studies, or PROSPERO for systematic reviews and meta-analyses), and is required for work to be published in several journals.

A first step of an analysis plan is the definition of the **study population** that it most appropriate to assess the main objective. **Inclusion criteria** (e.g. 'age \geq 50 years') and **exclusion criteria** (i.e. 'history of CRC') should unambiguously characterize the study population. Keep in mind that your findings will be only generalizable to a population with similar inclusion and exclusion criteria. Inclusion periods and **follow-up** should be clearly stated.

The term **outcome** refers to the measured variable (e.g. 'CRC mortality'), whereas the **endpoint** refers to its quantitative expression (for instance 'CRC-specific mortality within 5 years after colonoscopy screening'). For some specific questions, you can use composite endpoints (a combination of individual endpoints), which will have methodological implications that are outside the scope of this manuscript. All other items collected, including treatment/exposure (e.g. 'colonoscopy') and covariates, should be defined, with all possible values stated.

The study design (i.e. randomized, interventional vs. observational and/or prospective vs. retrospective), should be clearly defined, but a detailed discussion is outside the scope of this manuscript. In observational studies (notably retrospective), which will be the vast majority of projects for a young trainee, handling **missing data** should be pre-planned. The absence of randomization also requires the investigator to **control for confounding**. Confounders are parameters associated with both treatment/exposure and the outcome, and only randomization can reasonably ensure an equal distribution of confounders over the study population. For instance, parameters associated with health-seeking behaviour (i.e. 'participation in colonoscopy screening') are also associated with a healthier lifestyle (e.g. 'no smoking', 'healthier diet' and 'lower body weight'). Several methods have been applied to control for confounding. Conventional methods include stratification, matching and multivariate regression, while newer techniques include propensity scoring and marginal structural models.

The choice of the test is determined by the structure of the endpoint variable (quantitative vs. qualitative), whether observations are paired or unpaired, and how the variables are distributed. For a survival analysis, specific statistical tools (notably Kaplan–Meier and Cox analyses) are available. Hypothesis and **statistical significance** testing according to the p -value is based on proof by contradiction. The p -value is calculated under the assumption of the absence of an effect: the null hypothesis (i.e. ‘no difference among patients with or without colonoscopy screening’). The p -value is the probability of measuring an effect at least as strong as what was observed, given that there is no true difference between groups. The null hypothesis is typically rejected if the p -value falls below an arbitrary value of $p < 0.05$. However, this is no proof that the study hypothesis is true. Vice versa, a p -value > 0.05 is no proof of the absence of an effect. Smaller p -values alone do not guarantee an appropriate methodology. In the discussion of the clinical relevance of the results, magnitudes of effects (i.e. ‘1 vs. 20% CRC mortality reduction’) and the implications for the patient should be considered.

Subgroup and sensitivity analyses should be pre-planned to assess the robustness of the results according to changes in all aspects of study methodology. For instance, when testing in defined patient subgroups (e.g. ‘women’ and ‘ > 80 years’) with a different definition of the outcome variable (‘mortality reduction within 2 vs. 5 years’), a different statistical test or an alternative approach towards missing data can be performed. No strict rules apply and common sense should be employed. Consistent results in a limited number of pre-defined sensitivity analyses increase confidence towards the results of a study.

A **power analysis** is recommended to estimate the number of patients necessary to answer the primary research question. The results of the power analysis indicate how many patients should be included to detect a statistically significant difference (e.g. $p < 0.05$) with a given statistical power, for instance 80 or 90%. Power refers to the probability of detecting an effect when it is present. For a meaningful study, the size of this effect should also be of practical significance.

Data collection

Data collection is mostly performed by the trainee, frequently in retrospective studies including a review of medical records. It should only start after the finalization of the analysis plan, in order to avoid multiple data collection rounds. The analysis plan should include comprehensive definitions and descriptions (qualitative or quantitative) of all collected variables. When data are summarized there is a risk of losing information;

thus, collection of raw data with summarizing in the analysis phase is preferable.

Datasets should be de-identified, in order to protect individual privacy. It is recommended that a unique anonymous identifier is created as a study number for each patient, and that these identifiers are used throughout all datasets.

Data management, i.e. formatting the data to fit the requirements of the statistical analysis, can be time-consuming. Thus, data should be collected in a that manner minimizes data management. First, coded data are easier to analyze than free text. Second, variable names should respect the general naming conventions of the most used statistical software programmes. A commonly valid name starts with a letter and consists of a maximum of 32 characters including letters, numbers and the underscore character.

We recommend data collection using a Clinical Data Management System (CDMS), which is nowadays based on electronic data capturing (EDC) systems. An EDC system should also be used in retrospective observational studies, even those with a small sample size, since it has several advantages compared to spreadsheets. EDC systems allow restricted access in case of multiple users, while spreadsheets have very limited permission controls. Constraints for data entry are also available in EDC systems, which will immediately identify data errors or omitted data. Conversely, the use of spreadsheets may require a lengthy process of checking for errors and consolidating data. The use of Excel or similar spreadsheet software, even though widely available, is discouraged for data entry since such software does not meet most of the criteria mentioned above. The choice of the EDC software solution depends on the local institution; contact the local information technology staff to find out the current license software agreements. Research Electronic Data Capture (REDCap) is a web-based EDC software solution developed by Vanderbilt University, which is widely used in the academic research community (<https://www.project-redcap.org>).

Collaboration with the methodologist–statistician

Unless your mentor has advanced skills in methodology and statistics, you need to collaborate with a methodologist and a statistician. The methodologist will help the clinician to design the study, while the statistician will perform the statistical analysis and produce data analysis reports. In practice, it is mostly the same individual, who may already have some clinical knowledge in the topic of interest after previous collaborations with your research group. Such collaboration is facilitated if the clinician’s skills in methodology and

statistics, and the statistician's clinical knowledge, is high. Involving the statistician–methodologist in the study design process can enhance the quality of the study. Furthermore, the statistician–methodologist can help design the database and improve data collection. However, the clinician researcher should still maintain the initiative and lead the project in accordance with two points: answering the research question and providing clinical relevance.

Conclusion

Planning statistical analysis based on a research hypothesis can be a difficult task for an unexperienced trainee. All research projects should start with a clear main objective, from which the analysis plan will be developed. A research project can fail if data are

poorly collected. All of the tips provided here will help the researcher clinician to have a fruitful collaboration with the methodologist–statistician. Remember that the first research project is always the most difficult one and that you will gain experience over the years. The journey is worth it!

Julien Kirchgesner^{1,2} and Benjamin Misselwitz³

¹*Department of Gastroenterology, Hôpital Saint-Antoine, Paris, France*

²*Institut Pierre Louis d'Epidémiologie et de Santé Publique, Sorbonne Université, Paris, France*

³*Department of Gastroenterology and Hepatology, Universitätsspital Zürich and Zurich University, Zurich, Switzerland*